

# Nonparametric Bayesian testing for monotonicity

James G. Scott, Thomas S. Shively <sup>\*</sup>  
and Stephen G. Walker <sup>†</sup>

April 12, 2013

## Abstract

This paper studies the problem of testing whether a function is monotone from a nonparametric Bayesian perspective. Two new families of tests are constructed. The first uses constrained smoothing splines, together with a hierarchical stochastic-process prior that explicitly controls the prior probability of monotonicity. The second uses regression splines, together with two proposals for the prior over the regression coefficients. The finite-sample performance of the tests is shown via simulation to improve upon existing frequentist and Bayesian methods. The asymptotic properties of the Bayes factor for comparing monotone versus non-monotone regression functions in a Gaussian model are also studied. Our results significantly extend those currently available, which chiefly focus on determining the dimension of a parametric linear model.

Keywords: Bayesian asymptotics; Model selection; Monotonicity; Regression splines; Smoothing splines

## 1 Introduction

Many authors have used Bayesian methods for estimating functions that are known to have certain shape restrictions, including positivity, monotonicity, and convexity. Examples of work in this area include Holmes and Heard (2003), Neelon and Dunson (2004), Dunson (2005), Shively et al. (2009), Shively et al. (2011), and Hannah and Dunson (2012a,b). But an important question unaddressed by this work is whether it is appropriate to impose a specific shape constraint. If it is, the resulting function estimates are often considerably better than those obtained using unconstrained methods (Cai and Dunson, 2007). Conversely, the inappropriate imposition of constraints will result in poor function estimates.

---

<sup>\*</sup>McCombs School of Business, University of Texas at Austin; 2110 Speedway B6500, Austin, Texas, USA 78712; james.scott@mcombs.utexas.edu; tom.shively@mcombs.utexas.edu

<sup>†</sup>School of Mathematics, Statistics, and Actuarial Science, University of Kent, Kent, U.K; s.g.walker@math.utexas.edu

This paper considers the use of nonparametric Bayesian methods to test for monotonicity. The approach has connections with two major areas of the Bayesian literature. First, we extend work on regression and smoothing splines (Smith and Kohn, 1996; Dimatteo et al., 2001; Panagiotelis and Smith, 2008; Shively et al., 2009), and dictionary expansions more generally (Clyde and Wolpert, 2007), adapting these tools in such a way that they become appropriate for hypothesis testing. Second, we extend work on default Bayesian model selection (Berger and Pericchi, 2001; George and Foster, 2000; Clyde and George, 2004; Girón et al., 2010) to the realm of nonparametric priors for functional data.

Our main theoretical contribution is to characterize the asymptotic properties of the Bayes factor for comparing monotone versus non-monotone mean regression functions in a normal model. To date, such asymptotics have been exclusively orientated to comparing normal linear models of differing dimension. See, for example, Girón et al. (2010). We generalize this work to the problem of comparing non-linear mean functions. We do so using a technique introduced in Walker and Hjort (2002) that combines the properties of maximum-likelihood estimators and Bayesian marginal likelihoods that involve the square roots of likelihood functions.

Our main methodological contribution is to construct two new families of priors that are appropriate in a Bayesian test for monotonicity. The first approach uses a monotone version of smoothing splines, together with a hierarchical stochastic-process prior that allows explicit control over the time at which the underlying function's derivative first becomes negative. The second approach constructs a family of tests using a regression spline model (Smith and Kohn, 1996) with a mixture of constrained multivariate distributions as the prior for the regression coefficients. We study two possible choices for this prior: constrained multivariate Gaussian distributions, and a constrained version of the multivariate non-local method-of-moments distribution (Johnson and Rossell, 2010). The mixing parameters in both priors can be adjusted to allow the user to set the prior probability of a monotone function.

These approaches work for a very general class of sampling models, including those for continuous, binary, and count data. They are most easily understood, however, in the case where data  $Y = \{y_1, \dots, y_n\}$ , observed at ordered points  $x \in \{x_1, \dots, x_n\}$ , are assumed to arise from a normal sampling model,  $(y_i \mid f_i) \sim N(f_i, \sigma^2)$ . Here  $f_i = f(x_i)$  denotes the value at  $x_i$  of an unobserved, real-valued stochastic process  $f(x)$ , with  $x \in [0, 1]$  and  $f$  in some function space  $\mathcal{F}$ . To test whether  $f(x)$  is non-decreasing, one must compare the evidence for  $H_0 : f \in \mathcal{F}_0$  against  $H_1 : f \in \mathcal{F}_1$ , where

$$\begin{aligned}\mathcal{F}_0 &= \{f \in \mathcal{F} : f(s_2) \geq f(s_1) \text{ for all pairs } s_2 \geq s_1\}, \\ \mathcal{F}_1 &= \{f \in \mathcal{F} : f(s_2) < f(s_1) \text{ for at least one pair } s_2 \geq s_1\}.\end{aligned}$$

Under the Bayesian approach to this problem, two important goals are in conflict. One goal is flexibility: we wish to make few assumptions about  $f(x)$ , so that the testing procedure may accommodate a wide class of functions. Applying this principle

naïvely would lead us to choose an encompassing function space  $\mathcal{F}$  with large support, and vague prior measures  $\Pi_0$  and  $\Pi_1$  over  $\mathcal{F}_0$  and  $\mathcal{F}_1$ , respectively. On the other hand, the Bayes factor for comparing  $H_0$  versus  $H_1$  is

$$\text{BF}(H_0 : H_1) = \frac{\int_{\mathcal{F}_0} \left\{ \prod_{i=1}^N \phi(y_i | f(x_i), \sigma^2) \right\} d\Pi_0(f)}{\int_{\mathcal{F}_1} \left\{ \prod_{i=1}^N \phi(y_i | f(x_i), \sigma^2) \right\} d\Pi_1(f)},$$

which is heavily influenced by the dispersion of  $\Pi_0$  and  $\Pi_1$  (Scott, 2009). This strongly contra-indicates the use of noninformative priors for model selection. In particular, improper priors may not be used, as this leaves the Bayes factor defined only up to an arbitrary multiplicative constant.

To balance these conflicting goals, we let  $\mathcal{F} = \mathcal{C}^1$ , the space of all continuously differentiable functions. Membership in  $\mathcal{C}^1$  is easily enforced by supposing that  $f(x) = \int_0^x g(s) ds$ , where  $g(s) \in \mathcal{C}$ , the space of real-valued stochastic processes with almost-surely continuous sample paths. The test may now be phrased in terms of  $H_0 : g(s) \in C^+$  versus  $H_1 : g(s) \in C^-$ , where

$$\begin{aligned} \mathcal{C}^+ &= \{g \in \mathcal{C} : g(s) > 0 \text{ for all } s \in [0, 1]\}, \\ \mathcal{C}^- &= \mathcal{C} \setminus \mathcal{C}^+. \end{aligned}$$

Attention is then focused on the choice of prior for  $g(s)$ . The closest proposal to ours is that of Dunson (2005), who used a Bonferroni-like method for controlling the overall probability of monotonicity via a semiparametric model akin to a variable-selection prior. There is also a large body of classical work on specification tests, many of which are appropriate for assessing monotonicity or more general aspects of functional form. Two examples of work in this area are the papers by Zheng (1996), who proposed a test of functional form based on the theory of U-statistics, and Bowman et al. (1998), who used a two-stage bootstrap test for monotonicity based on the idea of a critical bandwidth parameter. Our simulation study finds that the new approaches proposed here outperform these methods, in that they have better power to detect departures from monotonicity at a fixed false-positive rate.

## 2 Constrained smoothing splines

A straightforward nonparametric Bayesian strategy to estimate a continuous function is to model the derivative  $g(x)$  as a scaled Wiener process prior, implying that the increments  $g(s_2) - g(s_1)$  are normally distributed with mean zero and variance  $\tau^2(s_2 - s_1)$ . But this is inappropriate for model selection, since it places unit prior probability on the hypothesis of non-monotonicity. To see this, recall that for every positive  $c$ , the sample path of a Wiener process almost surely takes both positive and negative values on  $(0, c)$ .

We therefore adapt this smoothing-spline approach to the testing problem. Let

$\xi = \inf_{s>0}\{s : g(s) = 0\}$  be a latent variable denoting the first downward crossing point of  $g(x)$ . This maps directly onto the hypotheses of interest:  $\xi \geq 1$  if and only if  $H_0$  is true, since the derivative  $g(s)$  will be strictly positive on the unit interval.

The introduction of  $\xi$  turns a hypothesis test for an infinite-dimensional object  $f$  into a one-dimensional estimation problem, via a hierarchical model for  $f(x)$ :

$$\begin{aligned} (y_i | f) &\sim N(f(x_i), \sigma^2), \quad f(x) = \int_0^x g(s) ds, \\ (g | \xi, \tau) &\sim \Pi(g | \xi, \tau), \quad \xi \sim p(\xi), \end{aligned} \quad (1)$$

with obvious modification in the non-Gaussian case. Here  $\Pi(g | \xi, \tau)$  denotes the conditional probability distribution of a Wiener process with scale parameter  $\tau$ , given  $\inf_{s>0}\{s : g(s) = 0\} = \xi$ . This differs from the one-dimensional Brownian bridge, in that  $g(x)$  is restricted to be positive over the interval  $(0, \xi)$ .

We refer to the overall approach as a constrained smoothing spline. To test  $H_0$  versus  $H_1$ , we compute the posterior distribution for  $\xi$ , and identify the posterior probability of  $H_0$  as  $\text{pr}(\xi > 1 | Y)$ . One default choice for  $p(\xi)$  is a standard log-normal distribution, but any proper prior with a median of 1 could be used.

As the data are observed on a discrete grid, it is necessary to characterize the increments of the above stochastic-process prior. The following proposition describes the distribution of these increments, here referred to as the fractional normal distribution.

**Proposition 1.** *Let  $\Pi(g | \xi, \tau)$  denote the conditional probability measure of a scaled Wiener process  $g(s)$  with scale parameter  $\tau$ , given the condition that the first downward crossing of  $g(x)$  occurs at  $x = \xi$ , where conditioning is meant in the sense of Doob's  $h$  transform. Let  $0 < s_1 < \xi$ , and define  $g_1 \equiv g(s_1)$  and  $g_0 \equiv g(0)$ . The conditional distribution  $p(g_1 | g_0, \xi, \tau)$  arising from  $\Pi(g | \xi, \tau)$  is a fractional normal distribution, denoted  $FN(g_1 | g_0, \xi, \tau)$ , with density*

$$p(g_1 | g_0, \xi, \tau) = \frac{\sqrt{2}e^{-m^2/2}}{mh^2\sqrt{\pi}} g_1 \sinh(mg_1/h) \exp\left\{-\frac{g_1^2}{2h^2}\right\}, \quad (2)$$

where  $h = \tau\{\xi u(1-u)\}^{1/2}$  and  $m = (g_0/\tau)\{(1-u)/(\xi u)\}^{1/2}$ .

*Proof.* The proof generalizes the argument of Chigansky and Klebaner (2008) to a scaled Wiener process, and follows their approach closely. Let  $V$  be a three-dimensional Brownian bridge with scale  $\tau$ , with  $V(0) = v$  and  $V(\xi) = 0$ , for some point  $v$  having Euclidean norm  $\|v\| = g_0$ . That is,

$$V(s) = v + W(s) - \frac{s}{\xi}\{W(\xi) + v\} \quad \text{for } s < \xi, \quad (3)$$

where  $W(s)$  is a Wiener process in three dimensions with scale parameter  $\tau$ . Observe that the random variable of interest is equal in distribution to the radial part of  $V$ ,

observed at  $s_1$ :

$$(g_1 \mid \xi, g_0) \stackrel{D}{=} \|V(s_1)\|.$$

Moreover, since  $\|V\|$  is independent of the starting point  $v$ , as long as  $\|v\| = g_0$ , one may choose  $v = (g_0, 0, 0)$  to simplify the calculations. From (3), this leads to

$$\|V(s_1)\| \stackrel{D}{=} \left\{ \left( z_1 \tau \sqrt{u(1-u)} \xi + g_1(1-u) \right)^2 + \xi u(1-u) \tau^2 z_2^2 + \xi u(1-u) \tau^2 z_3^2 \right\}^{1/2}, \quad (4)$$

where  $u = (\xi - s_1)/(\xi - s_0)$ , and where  $(z_1, z_2, z_3)$  are independent, standard-normal draws. Equivalently,

$$\|V(s_1)\| \stackrel{D}{=} h \{ (z_1 + m)^2 + z_2^2 + z_3^2 \}^{1/2},$$

with  $m$  and  $h$  defined as above. The density  $z_2^2 + z_3^2$  is that of an exponential distribution with rate  $1/2$ . Meanwhile, the density of the term  $\eta = (z_1 + m)^2$  may be evaluated directly using the fact that  $p(\eta) = \frac{d}{du} P(\eta < u)$  and differentiating under the integral sign. After a simple change of variables, the density of  $\theta = \{ (z_1 + m)^2 + z_2^2 + z_3^2 \}^{1/2}$  may be computed via convolution as

$$f_c(\theta) = \frac{\sqrt{2}e^{-m^2/2}}{m\sqrt{\pi}} \theta \sinh(m\theta) \exp(-\theta^2/2).$$

The result follows from the fact that the density of  $g_1 = h\theta$  is  $p(g_1) = f_c(\theta/h)/h$ .  $\square$

Proposition 1 characterizes the distribution of the increments of  $g(x)$  for a given  $\xi$ , and leads to an efficient sequential Monte Carlo algorithm for fitting the model. The details of the algorithm are in the appendix.

### 3 Constrained regression splines

This section develops a second family of tests for monotonicity using a regression spline model with different prior distributions for the regression coefficients. A finitely parametrized approximation to the function  $f(x)$  is given by the spline expansion

$$f_m(x) = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 (x - \tilde{x}_1)_+^2 + \cdots + \beta_{m+2} (x - \tilde{x}_m)_+^2, \quad (5)$$

where the  $\tilde{x}_j$  are the ordered knot locations, and  $z_+$  indicates the positive part of  $z$ . Throughout this section we will assume the model  $y_i = f_m(x_i) + \epsilon_i$  for some choice of  $m$ . For notational convenience let  $\tilde{x}_0 = 0$  and  $\tilde{x}_{m+1} = 1$ . Clearly the first derivative  $f'_m(x)$  is linear between each pair of knots. Therefore, if  $f'_m(\tilde{x}_j) \geq 0$  and  $f'_m(\tilde{x}_{j+1}) \geq 0$ , then  $f'_m(x) \geq 0$  on  $[\tilde{x}_j, \tilde{x}_{j+1}]$ . If this condition holds for all  $j = 0, \dots, m$ , then  $f_m(x)$  is monotone on  $[0, 1]$ .

To develop a test for monotonicity based on regression splines, we follow Smith

and Kohn (1996) and Shively et al. (2009), using Bayesian variable selection to choose the location of the knots from a set of  $m$  pre-specified values. Using all  $m$  knots, (5) may be re-written in matrix notation as  $y = \alpha 1 + X\beta + \epsilon$ , where  $y$  is the  $n \times 1$  vector of observations,  $1$  is a vector of ones,  $X$  is an  $n \times (m + 2)$  design matrix, and  $\beta$  is the vector of spline coefficients. Let  $\iota$  be a vector of indicator variables whose  $j$ th element takes the value 0 if  $\beta_j = 0$ , and 1 otherwise. Let  $\beta_\iota$  consist of the elements of  $\beta$  corresponding to those elements of  $\iota$  that are equal to one, and let  $p = |\iota|$  denote the number of nonzero entries in  $\iota$ . Shively et al. (2009) derive the constraints on  $\beta_\iota$  that ensure the monotonicity of  $f_m(x)$  for a given  $\iota$ . Specifically,  $f_m(x)$  is monotone whenever  $L_\iota \beta_\iota \equiv \gamma_\iota \geq 0$ , where  $L_\iota$  is a known lower-triangular matrix that depends on  $\iota$  and the  $\tilde{x}_j$ 's. Checking whether any element of  $\gamma_\iota$  is negative, and controlling the prior probability of such a constraint violation across all possible values of  $\iota$ , are the basis of our test for monotonicity. The matrix  $L_\iota$  therefore plays an important role in the prior distribution for  $\beta_\iota$ .

Given  $\iota$ , the  $\beta_\iota$  space is divided into  $2^p$  disjoint regions denoted  $R_\iota^{(1)}, \dots, R_\iota^{(2^p)}$ , with each region defined by a different combination of signs of  $f'_m(x)$  at each of the included knots. Without loss of generality we may let  $R_\iota^{(1)}$  denote the region where the derivative is non-negative at each of the included knots, in which case  $f'_m(x) \geq 0$  for all  $x \in [0, 1]$ . For a specific  $\iota$  and prior  $p(\beta_\iota)$ , one may compute the prior probability  $\text{pr}(\beta_\iota \in R_\iota^{(1)} \mid \iota)$ , which is identical to  $\text{pr}\{f'_m(x) \geq 0 \text{ for all } x \in [0, 1] \mid \iota\}$ .

The key feature of our approach is that we are able to specify priors on  $\iota$  and  $\beta_\iota$  that allow explicit control over the prior probability of a monotone function. Given these priors, we compute  $\text{pr}\{f'_m(x) \geq 0 \text{ for all } x \in [0, 1] \mid y\}$ , the posterior probability of a monotone function with  $\iota$  and  $\beta_\iota$  marginalized out. The function is declared to be monotone if this probability exceeds a specified threshold.

We now describe the overall approach for constructing  $p(\iota)$  and  $p(\beta_\iota)$ . The  $\iota_j$  are assumed to be independent with  $\text{pr}(\iota_j = 0) = p_j$ . Given  $\iota$  and the error variance  $\sigma^2$ , the prior for  $\beta_\iota$  is a discrete mixture of  $2^p$  multivariate distributions,

$$\beta_\iota \sim \sum_{d=1}^{2^p} q_d \Pi_d,$$

where  $\Pi_d$  is constrained to have support on  $R_\iota^{(d)}$ . Within each component of the mixture, there is a fixed combination of signs of  $f'_m(x)$  at each of the included knots. As  $R_\iota^{(1)}$  corresponds to the region where  $f'_m(x) \geq 0$  for all  $x$ ,  $q_1$  controls the prior probability of monotonicity.

Two specific choices for  $\Pi_d$  are considered: one based on the multivariate Gaussian distribution, and the other based on the multivariate method-of-moments distribution described by Johnson and Rossell (2010). Both priors involve the constraint matrix  $L_\iota$  in order to ensure that the necessary integrals are analytically feasible. This is analogous to the use of  $g$ -priors in ordinary linear regression. In the appendix these priors are constructed in detail, and a Markov-chain sampling algorithm is presented

for sampling from the joint posterior distribution over all model parameters.

## 4 Asymptotic properties of Bayes factors

### 4.1 Independent and identically distributed models

This section develops the asymptotic properties of the Bayes factor in a test for monotonicity. This analysis requires a new approach that significantly extends previous methods used to study Bayes factors in the context of parametric linear models.

To introduce the new approach, we first consider the case of independent and identically distributed observations. Suppose two Bayesian models,  $M_1$  and  $M_2$ , are to be studied and compared via the use of a Bayes factor:

$$B_{12} = \frac{\int_{\mathbb{F}_1} \prod_{i=1}^n f_1(x_i | \theta_1) \pi_1(d\theta_1)}{\int_{\mathbb{F}_2} \prod_{i=1}^n f_2(x_i | \theta_2) \pi_2(d\theta_2)}. \quad (6)$$

In the following,  $d_K(f, g) = \int f \log(f/g)$  is the Kullback–Leibler divergence between  $f$  and  $g$ , and

$$d_H(f, g) = \left\{ \int \left( \sqrt{f} - \sqrt{g} \right)^2 \right\}^{1/2}$$

is the Hellinger distance between  $f$  and  $g$ , which is bounded by 2.

**Theorem 2.** *Suppose that the data  $x_1, \dots, x_n$  are assumed to arise from some true density  $f_0(x)$ , and that models  $M_1$  and  $M_2$  are to be compared, where  $M_j = \{f_j(x | \theta_j), \pi_j(\theta_j), \theta_j \in \mathbb{F}_j\}$ . First, suppose that the true  $f_0(x)$  is in the Kullback–Leibler support of  $\pi_2$ . That is, for all  $\epsilon > 0$ ,*

$$\pi_2[\theta_2 : d_K\{f_0(\cdot), f_2(\cdot | \theta_2)\} < \epsilon] > 0. \quad (7)$$

*Second, suppose that for all sufficiently large  $n$  and for any  $c > 0$ , the following bound holds almost surely under  $f_0(x)$ :*

$$\sup_{\theta_1 \in \mathbb{F}_1} \prod_{i=1}^n \frac{f_1(x_i | \theta_1)}{f_0(x_i)} < e^{nc}. \quad (8)$$

*Finally, suppose that*

$$\inf_{\theta_1 \in \mathbb{F}_1} d_H\{f_1(\cdot | \theta_1), f_0(\cdot)\} > 0. \quad (9)$$

*Then  $B_{12} \rightarrow 0$  almost surely under  $f_0$ .*

*Proof.* To prove the result, consider the denominator of  $B_{12}$  in Equation (6) with a factor introduced, which is also introduced to the numerator, and so the factor cancels

out:

$$I_{n2} = \int_{\mathbb{F}_2} \prod_{i=1}^n \frac{f_2(x_i | \theta_2)}{f_0(x_i)} \pi_2(d\theta_2).$$

It is well known that with the first condition of the theorem,  $I_{n2} > e^{-n\tau}$  almost surely for all large  $n$  and for any  $\tau > 0$ . See, for example, Schwartz (1965).

Now consider the numerator with the same factor introduced:

$$I_{n1} = \int_{\mathbb{F}_1} \prod_{i=1}^n \frac{f_1(x_i | \theta_1)}{f_0(x_i)} \pi_1(d\theta_1).$$

We can write this with an upper bound as follows:

$$I_{n1} \leq \left\{ \sup_{\theta_1 \in \mathbb{F}_1} \prod_{i=1}^n \frac{f_1(x_i | \theta_1)}{f_0(x_i)} \right\}^{1/2} \int_{\mathbb{F}_1} \left\{ \prod_{i=1}^n \frac{f_1(x_i | \theta_1)}{f_0(x_i)} \right\}^{1/2} \pi_1(d\theta_1).$$

The second condition will ensure that the first term remains bounded. The second term, labelled as  $J_{n1}$ , has expectation

$$E(J_{n1}) \leq \int_{\mathbb{F}_1} \left[ 1 - \frac{1}{2} d_H^2 \{f(\cdot | \theta_1), f_0(\cdot)\} \right]^n \pi_1(d\theta_1).$$

Hence, with the third condition, for some  $\eta > 0$ ,  $E(J_{n1}) < e^{-n\eta}$  and thus  $J_{n1} < e^{-n\delta}$  almost surely for all large  $n$  and for some  $\delta > 0$ .

Putting these together, we have

$$B_{12} < \exp \left\{ -n(\delta - \frac{1}{2}c - \tau) \right\} \quad \text{almost surely for all large } n \text{ for any } c, \tau > 0.$$

Choose  $\frac{1}{2}c + \tau < \delta$  to obtain the desired result.  $\square$

Using a similar argument, it is also possible to show that  $B_{12} \rightarrow \infty$  almost surely if the following three conditions hold. First, the true  $f_0(x)$  is in the Kullback–Leibler support of  $\pi_1$ ; that is, for all  $\epsilon > 0$ ,

$$\pi_1[\theta_1 : d_K\{f_0(\cdot), f_1(\cdot | \theta_1)\} < \epsilon] > 0.$$

Second,

$$\sup_{\theta_2 \in \mathbb{F}_2} \prod_{i=1}^n \frac{f_2(x_i | \theta_2)}{f_0(x_i)} < e^{nc} \quad \text{almost surely for all large } n \text{ for any } c > 0.$$

Finally,

$$\inf_{\theta_2 \in \mathbb{F}_2} d_H\{f_2(\cdot | \theta_2), f_0(\cdot)\} > 0.$$



## 4.2 Regression models: monotone vs. non-monotone

We will now adapt this result to examine the regression spline model introduced previously. Suppose that  $M_1$  is a normal model with monotone mean function,

$$M_1 = \{g_1(y \mid x, \theta_1) = N(y \mid f_1(x), \sigma_1^2), \pi_1(f_1, \sigma_1^2)\},$$

and  $M_2$  is a normal model with non-monotone mean function,

$$M_2 = \{g_2(y \mid x, \theta_2) = N(y \mid f_2(x), \sigma_2^2), \pi_2(f_2, \sigma_2^2)\}.$$

Assume the  $x_i$  are sampled from distribution  $Q$  with support on  $(0, 1)$ . Let  $f$  be a regression spline function, which can be monotone or non-monotone, and define the sieve

$$S'_n = \left\{ f \in C : f \text{ has knot points at the } (x_i) \text{ and } \int |f'(x)|^2 dx < \lambda_n \right\}$$

for some  $\lambda_n \uparrow +\infty$ , where  $C$  is the space of continuous, piecewise linear functions on  $(0, 1)$ .

Finally, define

$$B_{12} = \frac{\int \prod_{i=1}^n N\{y_i \mid f_1(x_i), \sigma_1^2\} \pi_1(d\theta_1)}{\int \prod_{i=1}^n N\{y_i \mid f_2(x_i), \sigma_2^2\} \pi_2(d\theta_2)},$$

where  $\theta_j = (f_j, \sigma_j)$ .

Our main result is stated below.

**Theorem 3.** *Assume that  $f_0$ , the true regression function is bounded and continuous on  $(0, 1)$  and, if monotone, has a derivative bounded away from 0. Suppose that  $\lambda_n = O(n^{1/4-\delta})$  for some  $\delta > 0$ ; that the  $L_2$  support of the prior for  $f$ , which may depend on the sample size, coincides with  $S'_n$ ; and that the support of the variances coincides with  $\sigma_1, \sigma_2 < \sigma_+ < \infty$ . Then, if the true model lies in  $M_1$ ,  $B_{12}$  diverges almost surely, whereas if the true model lies in  $M_2$ , then  $B_{12} \rightarrow 0$  almost surely.*

The full proof is somewhat technical and is provided in the Supplementary material. The essential ideas are conveyed by the proof of Theorem 1, while many of the technical details draw on the method of sieves studied by Geman and Hwang (1982). We now briefly describe the connection between the theorem and our regression-spline approach, and the way the conditions of the theorem correspond to the three conditions of Theorem 1.

The prior for  $f$ , say  $\pi^{(n)}$ , will depend on the sampled  $x_i$  and actually have as support the functions in  $S'_n$ . In order to ensure the true model is in the Kullback-Leibler support of the prior for all large  $n$ , we assume the true variance is bounded by  $\sigma_+$ , and that for all large  $n$ ,  $f_0$  is in the  $L_2$  support of the prior. That is, for all

$\epsilon > 0$ , and sufficiently large  $n$ ,

$$\pi^{(n)} \left\{ f : \int (f_0(x) - f(x))^2 Q(dx) < \epsilon \right\} > 0.$$

In the proof of Theorem 2 we need the condition that for some  $t > 0$ ,

$$\int e^{t|y|} F_Y(y) dy < \infty,$$

where  $F_Y$  is the true marginal distribution of the  $y$ . This condition on  $F_Y$  and  $\lambda_n$  is to ensure that the sieve maximum likelihood estimator of  $f_0$  exists and converges appropriately for our purposes. Here

$$F_Y(dy) = \int N(dy|f_0(x), \sigma_0^2) Q(dx),$$

and hence the condition on  $F_Y$  will be satisfied if

$$\int \exp \{f_0^2(x)\} Q(dx) < \infty.$$

This holds when  $f_0$  is bounded, which we assume in Theorem 2.

When  $f_0$  is monotone, we need to assume that  $f'_0$  is bounded away from 0 to ensure that the probability a non-monotone function can get arbitrarily close to it is sufficiently small. We can not ensure this sufficient smallness in probability if  $f_0$  is flat. The idea is that if the non-monotone function has probability 1/2 of increasing or decreasing at every  $x_i$  point, then the probability it increases  $n - 1$  times but decreases once is exponentially small. On the other hand, it is not possible for a monotone function to get arbitrarily close to a non-monotone function.

## 5 Experiments

### 5.1 Description of simulations

This section reports the results of an extensive simulation study benchmarking the proposed methods against three alternatives that previously appeared in the literature. The first two are well-known in the classical literature: the U test from Zheng (1996), and the bootstrap-based test from Bowman et al. (1998). Our third benchmark was the method of Bayesian Bonferroni correction proposed by Dunson (2005), where each increment of  $f$  is as

$$\begin{aligned} (y_i | f_i, \sigma^2) &\sim N(f_i, \sigma^2), \quad f_i = f_{i-1} + \delta_i, \\ (\delta_i | w, \tau^2) &\sim wN^+(\delta_i | 0, \tau^2) + (1 - w)N(\delta_i | 0, \tau^2), \end{aligned}$$

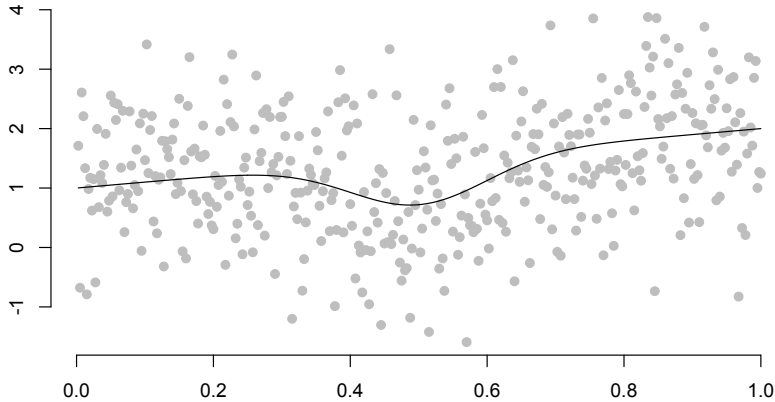


Figure 1: The Bowman et. al. test function with  $a = 0.78$ .

where  $w$  is the mixing probability and  $N^+$  indicates a normal distribution truncated below at zero. By analogy with Bonferroni correction, one then chooses  $w$  as a function of the sample size such that the event  $\{\delta_i > 0 \text{ for all } i\}$  has prior probability  $1/2$ . This model can be fit straightforwardly via Gibbs sampling.

Our simulations used the model  $y_i = f(x_i) + \epsilon_i$ , with  $n = 400$  equally spaced  $x$  values on  $(0, 1]$ , and the  $\epsilon_i$  independent and identically distributed  $N(0, 1)$  random variables. We generated 100 data sets for each of the 13 functions listed in the left-most column of Table 1. Five of these functions are monotone, but either flat or very nearly flat compared to the chosen error variance of  $\sigma^2 = 1$ . The remaining 8 functions are non-monotone, but only just. The goal of the experiment was to assess the performance of all methods across a range of situations with low signal-to-noise ratios. The first 11 test functions involving simple linear, quadratic, and exponential forms. The last two test functions are of the form  $f(x) = 1 + x - a \exp\{-0.5(x - 0.5)^2/0.1^2\}$ , which is used by Bowman et al. (1998) in their simulation study. Setting  $a = 0.78$  gives a decrease of 0.5 from the maximum of the function on the interval  $(0, 0.5)$  to the minimum of the function, while setting  $a = 1.05$  gives a corresponding decrease of 0.75. See Figure 1 for the case where  $a = 0.78$ , along with one simulated data set. In Table 1 we refer to these functions as BJG, for Bowman, Jones, and Gijbels.

For the two regression spline-based tests developed in Section 3, we used  $m = 9$  equally spaced knots, and set  $c = 5$ ,  $p_j = \text{pr}(\iota_j = 0) = 0.75$ , and  $q_1 = \text{pr}(f'_m(x) \geq 0 \mid \iota) = 0.5$ . The simulation results indicate that these values give procedures with good small-sample classification properties for a wide range of functions. The MCMC sampling scheme for each procedure was run for a burn-in period of 20,000 iterations and a sampling period of 100,000 iterations.

For the constrained smoothing-spline approach, we used a particle filter with 50,000 particles. We assumed a Gamma(2,2) prior on  $\tau^2$ , the variance of the underlying Wiener process for  $f'(x)$ . For the crossing time  $\xi$ , we assumed a standard log-normal prior, which has 50% of its prior probability on the unit interval. The

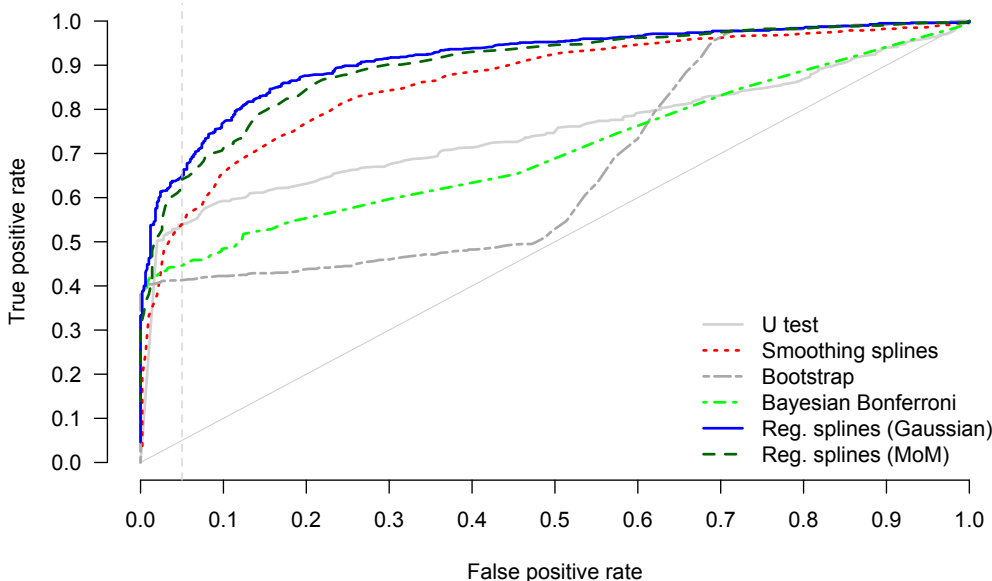


Figure 2: Receiver operating characteristic curves for the different methods, evaluated across all 1300 data sets simulated under the functions listed in Table 1. The diagonal line is the line of unit slope, while the vertical dashed line is at  $\alpha = 0.05$ .

plug-in estimator described in Section 2 was used for the error variance  $\sigma^2$ .

## 5.2 Results

The receiver-operating characteristic curve in Figure 2 provides a common scale to judge the overall performance of the various methods across the range of signal classes considered here. To calculate this curve, observe that all tests can be phrased in terms of a test statistic  $t$  and a critical value  $t^*$ , where the null hypothesis of monotonicity is rejected if  $t < t^*$ . For all Bayesian methods, we used the Bayes factor against monotonicity as the test statistic. For the bootstrap and U-tests, we used the  $p$ -value under the hypothesis of monotonicity as the test statistic. At a given critical value  $t^*$ , the realized false positive rate and false negative rate may be calculated across all 1300 tests. Thus we are calculating an average performance across the 13 different functions, some of which are monotone and some of which are not. The curves are then traced out as one varies the critical value independently for each method.

As Figure 2 shows, the Bayesian methods proposed here exhibit the best overall frequentist properties for functions among those considered. At an  $\alpha$  level of 0.05, the smoothing-spline approach has similar power to the U test, and noticeably better power at all higher  $\alpha$  levels. The regression spline method had uniformly higher power than the other methods across all  $\alpha$  levels, regardless of whether the Gaussian-

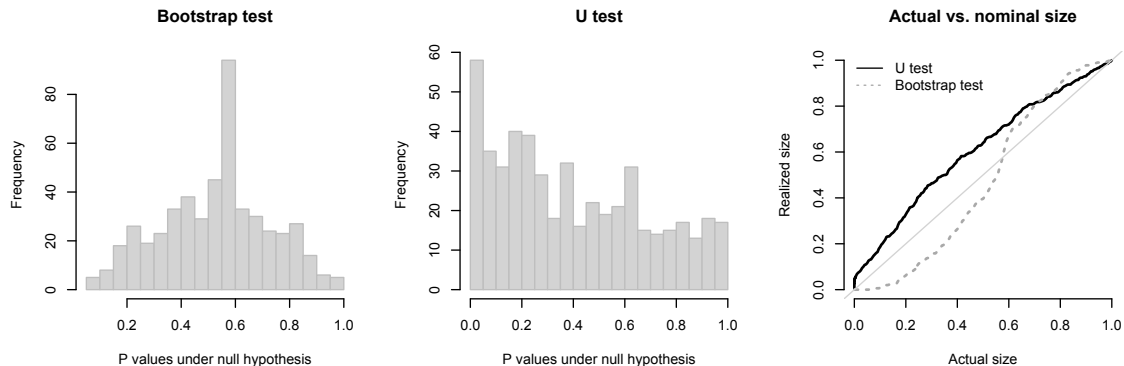


Figure 3: The empirical sampling distribution of  $p$ -values under the null hypothesis for the bootstrap test, left panel, and U test, middle panel. In neither case is the distribution uniform. Right: actual versus nominal size

or moment-based prior is used. The Bayesian-Bonferroni method and the U tests behaved similarly overall, with the U- test giving slightly better power for most  $\alpha$  levels. The bootstrap-based test performed the poorest, having the least overall area under the receiver-operating characteristic curve.

Figure 3 shows the empirical distribution under the null hypothesis of the  $p$ -values calculated using the bootstrap test and U test. There were 500 such cases in our simulation where the null hypothesis was true. These 500  $p$ -values were not uniformly distributed on  $(0, 1)$  under either method, although their distribution under the U test was much closer to uniform. These departures from uniformity account for the phenomenon observed in the right panel of Figure 3, which shows the realized false-positive rate for the bootstrap and U tests as a function of the nominal size of the rejection region. In practical terms, this plot suggests that, when using the U test, one may need to choose a rejection region of nominal size  $\alpha > 0.05$  in order to produce an actual false-positive rate of 0.05 under the null.

Table 1 shows more detail about the results of the simulation study for each function. We calibrated the rejection region for each method to have a realized  $\alpha$  level of 0.05 across all functions, indicated by the vertical dashed line in Figure 2. This calibration is necessary to make a fair comparison across the different functions, and as shown in Figure 3, is not the same as the nominal  $\alpha$  level for the bootstrap and U tests. We then tallied the number of correct classifications for each function at this calibrated critical value. A few interesting differences emerge. For example, the U test has superior power for the Bowman et. al. functions compared to the Bayesian methods, while the smoothing-spline approach does worse for the functions with dips near  $x = 1$ .

One further aspect of the Bayesian method is that it can be used to simultaneously test global and local hypotheses about the behavior of the function, without invoking any concerns about multiplicity. This becomes clear when examining a spe-

Table 1: The entries in the table show the percentage of times in 100 Monte Carlo simulations that the given function was classified correctly, with each method’s rejection region calibrated to have a size of  $\alpha = 0.05$  across all 13 functions. RS: constrained regression splines (Gaussian and MoM priors). SS: constrained smoothing spline with Wiener-process prior. BB: Bayesian Bonferroni. Boot: bootstrap-based test from Bowman et al. (1998). U: nonparametric U-test from Zheng (1996).

Function	Monotone?	Method					
		U	Boot	SS	BB	RS (Gauss)	RS (MoM)
$-0.292 \exp(x^3)$	No	15	0	91	19	92	85
$-0.146 \exp(x^3)$	No	0	0	48	0	44	33
$0.146 \exp(x^3)$	Yes	98	98	97	98	97	97
$0.292 \exp(x^3)$	Yes	91	87	98	93	100	100
Flat	Yes	100	99	85	99	85	90
$1 + 4.0(x - 0.75)^2$	No	94	81	15	96	54	58
$1 + 8.0(x - 0.75)^2$	No	97	92	36	99	70	74
$-0.50x$	No	27	0	91	28	92	86
$-0.25x$	No	2	0	58	4	46	34
$0.25x$	Yes	99	96	99	95	97	95
$0.50x$	Yes	88	94	100	91	97	94
BJG, $a = 1.05$	No	99	92	60	71	87	89
BJG, $a = 0.78$	No	98	65	20	29	31	35

cific case. In Figure 4 we see two examples of data sets in our simulation study: one where  $f(t) = 1 + 8.0(t - 0.75)^2$ , which is non-monotone; and the other where  $f(t) = 0.5t$ , which is monotone. For each of these two data sets, the  $U$ -test rejected the null hypothesis of monotonicity at the  $\alpha = 0.05$  level. The Bayesian test using constrained Brownian motion also favors the hypothesis of non-monotonicity for each case: posterior probability 0.98 for the function in the top pane, and 0.74 for the one in bottom pane.

The problem comes when we try to identify the range of values where  $f$  exhibits likely non-monotonic behavior. The  $U$ -test, intuitively, asks whether an unconstrained estimate of  $f(t)$  significantly improves the residual sum of squares, versus a constrained estimate. In this sense it behaves like an  $F$ -test in a classical analysis of variance problem. It is an omnibus test, and leads to a large multiple-comparison problem if one wishes to test for specific local features of the data set that yielded a rejection from the omnibus test. In contrast, the posterior probability associated with any such local question, such as whether the function decreases between 0.6 and 0.8, arises naturally from the joint posterior distribution under the Bayesian model. One may ask an unlimited number of such questions of the posterior distribution, without posing any multiplicity concerns.

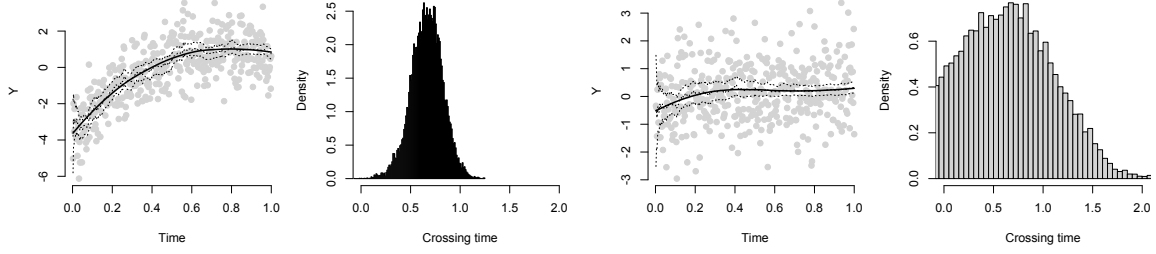


Figure 4: The left two panels show results for a data set where  $E(y) = 1 + 8.0(t - 0.75)^2$ . The first panel shows the data as grey dots, the filtered mean and 95% filtered confidence interval for  $f$  as dotted lines, and the smoothed mean as a solid line. The second panel shows the histogram of draws from the posterior distribution of the first crossing time  $\xi$ . The right two panels show the same two figures for a data set where  $E(y) = 0.5t$ .

## A Further detail on constrained smoothing splines

Proposition 1 in the main manuscript characterizes the distribution of the increments of the stochastic process  $g(s)$  for a particular value of the crossing time  $\xi$ . It naturally suggests a sequential Monte Carlo algorithm for model-fitting. Let  $\Delta_j = t_{j+1} - t_j$  denote the  $j$ th time increment, and approximate the value of  $f$  at time  $t_i$  as  $f(t_i) \approx \sum_{j < i} g(t_j) \Delta_j$ , assuming that the time bins are sufficiently small such that  $g(t_j) \Delta_j$  is a good approximation of  $\int_{t_j}^{t_{j+1}} g(s) ds$ .

Subject to this approximation,

$$(y_i | f_i, \sigma^2) \sim N(f_i, \sigma^2) \quad (10)$$

$$f_i = f_{i-1} + \Delta_{i-1} g_{i-1} \quad (11)$$

$$(g_i | \tau, \xi, g_{i-1}) \sim \begin{cases} \text{FN}(g_i | g_{i-1}, \xi, \tau) & \text{if } t_j < \xi, \\ N(g_i | 0, \tau^2(t_j - \xi)) & \text{if } t_{j-1} < \xi < t_j, \\ N(g_i | g_{i-1}, \tau^2(t_j - t_{j-1})) & \text{if } t_{j-1} > \xi, \end{cases} \quad (12)$$

recalling that FN denotes the fractional normal distribution from Proposition 1. The model is characterized by the unobserved state vector  $g = (g_1, \dots, g_N)^T$ , and the three unknown parameters  $\xi$ ,  $\tau$ , and  $\sigma^2$ . To fit it, we use the particle-filtering algorithm from Liu and West (2001), which is a modification of sequential importance resampling. The idea is to introduce a particle approximation to the time- $t$  filtered distribution of the state  $g_t$  and unknown parameters  $\tau$  and  $\xi$ . The problem of particle decay is handled by propagating each particle forward with a small jitter added to  $\xi$  and  $\tau$ , drawn from Gaussian kernels centered at the current values. This would ordinarily result in a density estimate that is over-dispersed compared to the truth. But Liu and West (2001) suggest a shrinkage correction that shifts the propagated values toward their overall sample mean, resulting in a particle approximation with the correct degree of dispersion.

With a moderate amount of data,  $\sigma^2$  can be estimated quite precisely. We use a simple plug-in estimate, derived from an unconstrained local-linear-regression esti-

mate of  $f$ , with the kernel bandwidth chosen by leave-one-out cross-validation. This produces an estimate for  $\sigma^2$  that is asymptotically efficient, regardless of whether the underlying function is monotone. We also experimented with the particle-learning approach of Carvalho et al. (2010), by tracking sufficient statistics for  $\sigma^2$  as part of the state vector. Only very small differences between these two approaches were observed, seemingly validating the simpler plug-in method. Code implementing the method is available from the authors.

## B Further details on constrained regression splines

### B.1 The model

This section develops two tests of monotonicity using a regression spline model with different prior distributions for the regression coefficients. Equation (5) from the main manuscript specifies a finitely parametrized approximation to the function  $f(x)$ , and introduces our notation for the problem. For the sake of clarity, we re-iterate this notation here. We represent the unknown function as

$$f_m(x) = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 (x - \tilde{x}_1)_+^2 + \cdots + \beta_{m+2} (x - \tilde{x}_m)_+^2,$$

where the  $\tilde{x}_j$  are the ordered knot locations, and  $z_+$  indicates the positive part of  $z$ . Using all  $m$  knots, this model may be re-written in matrix notation as  $y = \alpha 1 + X\beta + \epsilon$ , where  $y$  is the  $n \times 1$  vector of observations,  $1$  is a vector of ones,  $X$  is an  $n \times (m+2)$  design matrix, and  $\beta$  is the vector of spline coefficients. Let  $\iota$  be a vector of indicator variables whose  $j$ th element takes the value 0 if  $\beta_j = 0$ , and 1 otherwise. Let  $\beta_\iota$  consist of the elements of  $\beta$  corresponding to those elements of  $\iota$  that are equal to one, and let  $p = |\iota|$  denote the number of nonzero entries in  $\iota$ . Shively et al. (2009) derive the constraints on  $\beta_\iota$  that ensure the monotonicity of  $f_m(x)$  for a given  $\iota$ . Specifically,  $f_m(x)$  is monotone whenever  $L_\iota \beta_\iota \equiv \gamma_\iota \geq 0$ , where  $L_\iota$  is a known lower-triangular matrix that depends on  $\iota$  and the  $\tilde{x}_j$ 's.

Given  $\iota$ , the  $\beta_\iota$  space is divided into  $2^p$  disjoint regions denoted  $R_\iota^{(1)}, \dots, R_\iota^{(2^p)}$ , with each region defined by a different combination of signs of the derivative  $f'_m(x)$  at each of the included knots. In general, if the sign of  $f'_m(x)$  at any included knot is negative, then the function is not monotone. Without loss of generality we may let  $R_\iota^{(1)}$  denote the region where the derivative is non-negative at each of the included knots, in which case  $f'_m(x) \geq 0$  for all  $x \in [0, 1]$ . For a specific  $\iota$  and prior  $p(\beta_\iota)$ , one may compute the prior probability  $\text{pr}(\beta_\iota \in R_\iota^{(1)} \mid \iota)$ , which is identical to  $\text{pr}\{f'_m(x) \geq 0 \text{ for all } x \in [0, 1] \mid \iota\}$ .

Given  $\iota$  and  $\sigma^2$ , the prior for  $\beta$  is a mixture of  $2^p$  distributions, with the  $d$ th mixture component  $\Pi_d$  constrained to have support on  $R_\iota^{(d)}$ . Two specific choices for  $\Pi_d$  are considered: one based on the multivariate Gaussian distribution, and the other based on the multivariate method-of-moments distribution described by Johnson and Rossell (2010). As discussed below and shown in the simulation experiment in Section 5 of the main manuscript, the two resulting tests have different small sample properties.



We first note that a multivariate Gaussian distribution for the regression coefficients is inappropriate for testing purposes, because for any given region  $R_\iota^{(d)}$ ,  $\text{pr}(\beta_\iota \in R_\iota^{(d)} \mid \iota)$  will be very small if there are a large number of knots in the model. Because  $R_\iota^{(1)}$  is identified with the region where  $f(x)$  is monotone, this means there is a vanishingly small prior probability on the null hypothesis as the number of knots increases.

We therefore take the following approach. For a given  $\iota$  not identically 0, let  $L_\iota$  be the lower-triangular matrix that defines the monotonicity constraint on  $\beta_\iota$ . Let  $c$  be a fixed scale factor. The Gaussian-based prior we use is

$$(\beta_\iota \mid \iota, \sigma^2) \sim \sum_{d=1}^{2^p} q_d \text{TN} \left\{ \beta_\iota \mid 0, c\sigma^2(L_\iota L_\iota^T)^{-1}, R_\iota^{(d)} \right\},$$

where  $\text{TN}(x \mid m, V, R)$  denotes the density function, evaluated at  $x$ , of the multivariate normal distribution with mean vector  $m$ , covariance matrix  $V$ , and truncation region  $R$ . The integral of the density function over each of the  $2^p$  mutually disjoint regions may be done analytically, due to the choice of covariance matrix. In particular,  $\text{pr}(\beta_\iota \in R_\iota^{(d)} \mid \iota) = q_d$ , so conditional on  $\iota$ ,  $q_1$  is the prior probability of a monotone function. A reasonable choice here is  $q_1 = 1/2$ , with all the remaining  $q_d$ 's being equal. A second alternative is to choose the  $q_d$  values so the prior probability that the derivative function crosses zero between any pair of knots is the same.

As a second choice of prior, we also consider the multivariate moment-based prior,

$$(\beta_\iota \mid \iota, \sigma^2) \sim C_\iota \sum_{d=1}^{2^p} q_d \beta_\iota^T \Sigma_\iota^{-1} \beta_\iota |\Sigma_\iota|^{-1/2} \exp \left\{ -\frac{1}{2} \beta_\iota^T \Sigma_\iota^{-1} \beta_\iota \right\},$$

where  $C_\iota$  is a known normalizing constant not depending on  $\beta$ , and where  $\Sigma_\iota = c\sigma^2(L_\iota L_\iota^T)^{-1}$  as before. The integral of this density function over each of the subregions can be done analytically to give  $\text{pr}(\beta_\iota \in R_\iota^{(d)} \mid \iota)$ . As in the parametric linear case considered by Johnson and Rossell (2010), using the moment-based prior gives a test for monotonicity with different properties than the test obtained using a Gaussian prior. These differences are illuminated in the simulation results reported in Section 5 of the main manuscript.

To complete the model for  $f_m(x)$  we must also specify priors for  $\alpha$ ,  $\sigma^2$  and  $\iota$ . The intercept  $\alpha$  is given a vague mean-zero normal prior with variance  $10^{10}$ , and the variance is given a flat prior on  $[0, 10^3]$ . Using vague priors on these parameters is acceptable, as they appear in all models under consideration, and there is no indeterminacy in the Bayes factor as a result. The prior for the knot-inclusion indicator  $\iota$  was discussed in the main manuscript; each element is given a Bernoulli prior with fixed probability  $p_j$ .

The prior probability of monotonicity may then be computed as

$$\text{pr}\{f'_m(x) \geq 0 \text{ for all } x \in [0, 1]\} = q_1 [1 - \text{pr}\{\iota = (0, \dots, 0)\}] + \text{pr}\{\iota = (0, \dots, 0)\},$$

since  $q_1 = \text{pr}\{f'_m(x) \geq 0 \text{ for all } x \in [0, 1] \mid \iota\}$ , and since a flat function with  $\iota = (0, \dots, 0)$  is also a monotone function.

## B.2 Sampling scheme for posterior inference

To construct our Markov-chain Monte Carlo sampling scheme, we use an alternative parametrization of a regression spline model proposed in Shively et al. (2009). For a given  $\iota$ , define  $W_\iota = X_\iota L_\iota^{-1}$  and  $\gamma_\iota = L_\iota \beta_\iota$ , where  $L_\iota$  is the constraint matrix defined previously, and where  $X_\iota$  consists of the columns of  $X$  corresponding to the nonzero entries in  $\iota$ . This allows us to rewrite the function as  $f_m(x) = \alpha 1 + W_\iota \gamma_\iota$ , and to identify the regions  $R_\iota^{(d)}$  as orthants in the transformed  $\gamma$  space. In particular, the region of monotonicity is the first orthant, where  $\gamma_\iota \geq 0$ . This greatly simplifies the sampling scheme.

Under this new parametrization, the Gaussian-based prior becomes

$$(\gamma_\iota \mid \iota, \sigma^2) \sim \sum_{d=1}^{2^p} q_d \text{TN} \{ \gamma_\iota \mid 0, c\sigma^2 I, R_\iota^{(d)} \} ,$$

while the moment-based prior becomes

$$(\gamma_\iota \mid \iota, \sigma^2) \sim C_\iota \sum_{d=1}^{2^p} q_d \gamma_\iota^T \gamma_\iota \exp \left\{ -\frac{1}{2c\sigma^2} \gamma_\iota^T \gamma_\iota \right\} .$$

Sampling  $\alpha$  and  $\sigma^2$  is straightforward, so the details of these steps are omitted. We now discuss the details for sampling  $\iota$  and  $\gamma_\iota$  under the multivariate method-of-moments prior, the details for the Gaussian-based prior being similar.

Let  $\iota_j$  denote the  $j$ th element of  $\iota$ , and  $\iota_{-j}$  denote the remaining  $j-1$  elements. Similarly, let  $\gamma_j$  denote the  $j$ th element of  $\gamma$ , and  $\gamma_{-j}$  the remaining elements. We sample each  $(\iota_j, \gamma_j)$  jointly, given  $(\iota_{-j}, \gamma_{-j})$ , the data, and all other model parameters. To keep notation simple, we let  $\Theta$  denote the complete set of other model parameters, including the entries in  $\iota$  and  $\gamma$  not being sampled.

We generate  $(\iota_j, \gamma_j \mid y, \Theta)$  by first generating  $(\iota_j \mid y, \Theta)$  marginalizing over  $\gamma_j$ , and then generating  $(\gamma_j \mid \iota_j, y, \Theta)$ .

Without loss of generality assume that we are updating  $\iota_1$ . To compute  $p(\iota_1 = 0 \mid y, \Theta)$ , observe that if at least one element of  $\iota_{-1}$  is nonzero, then

$$p(\iota_1 = 0 \mid y, \Theta) = C p(y \mid \iota_1 = 0, \Theta) p(\gamma_{-1} \mid \iota_1 = 0, \iota_{-1}) p(\iota_1 = 0) , \quad (13)$$

where  $C$  is a constant. The second term on the right-hand side is

$$p(\gamma_{-1} \mid \iota_1 = 0, \iota_{-1}) = \tilde{q} 2^s (cs)^{-1} \gamma_{-1}^T \gamma_{-1} (2\pi c)^{-s/2} \exp \left\{ -\frac{1}{2c} \gamma_{-1}^T \gamma_{-1} \right\} ,$$

where  $s = |\iota_{-1}|$  is the number of nonzero elements in  $\iota_{-1}$ ; and where  $\tilde{q} = q_1$  if all elements of  $\gamma_{-1}$  are positive, and  $(1 - q_1)/(2^s - 1)$  otherwise. If all elements of  $\iota_{-1}$  are zero, then the same representation holds with  $p(\gamma_{-1} \mid \iota_1 = 0, \iota_{-1})$  set to one.

To compute  $p(\iota_1 = 1 \mid y, \Theta)$ , note that

$$p(\iota_1 = 1 \mid y, \Theta) = C \left\{ \int_{\mathbb{R}} p(y \mid \iota_1 = 1, \gamma_1, \Theta) p(\gamma_1, \gamma_{-1} \mid \iota_1 = 1, \iota_{-1}) d\gamma_1 \right\} p(\iota_1 = 1) , \quad (14)$$

where  $C$  is the same constant appearing in Equation (13). Let  $\delta_1 = (y - \alpha 1 - W_{-1}\gamma_{-1}) - w_1\gamma_1$ , with  $w_1$  representing the first column of  $W_{1,\iota_{-1}}$ , and  $W_{-1}$  the remaining columns. The first term in the integrand in (14) may be written as

$$p(y \mid \iota_1 = 1, \Theta) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \delta_1^T \delta \right\}.$$

Also,  $p(\gamma_1, \gamma_{-1} \mid \iota_1 = 1, \iota_{-1})$  is given by

$$a(\gamma_1) 2^r (rc)^{-1} (\gamma_1^2 + \gamma_{-1}^T \gamma_{-1}) (2\pi c)^{-r/2} \exp \left\{ -\frac{1}{2c} (\gamma_1^2 + \gamma_{-1}^T \gamma_{-1}) \right\}, \quad (15)$$

where  $r = s + 1$ ,  $a(\gamma_1) = \tilde{q}$  if  $\gamma_1 > 0$ , and  $(1 - q_1)/(2^r - 1)$  if  $\gamma_1 < 0$ . Recall that  $\tilde{q} = q_1$  if all elements of  $\gamma_{-1}$  are positive, and  $(1 - q_1)/(2^r - 1)$  otherwise.

Let  $\tilde{y} = y - \alpha 1 - W_{-1}\gamma_{-1}$ , and let

$$d = 2^r (cr)^{-1} c^{-1/2} (2\pi c)^{-s/2} (2\pi\sigma^2)^{-n/2} \exp \left( -\frac{1}{2c} \gamma_{-1}^T \gamma_{-1} \right) \left( \frac{w_1^T w_1}{\sigma^2} + c^{-1} \right)^{-1/2} p(\iota_1 = 1).$$

Then (14) can be written as

$$\begin{aligned} p(\iota_1 = 1 \mid y, \Theta) &= Cd \exp \left[ -\frac{1}{2\sigma^2} \left\{ \frac{\tilde{y}^T \tilde{y} - (w_1^T \tilde{y})^2}{w_1^T w_1 + \sigma^2/c} \right\} \right] \\ &\times \left\{ \frac{1 - q_1}{2^r - 1} \int_{-\infty}^0 h(\gamma_1) d\gamma_1 + \tilde{q} \int_0^{\infty} h(\gamma_1) d\gamma_1 \right\}, \end{aligned} \quad (16)$$

where

$$\begin{aligned} h(\gamma_1) &= (\gamma_1^2 + \gamma_{-1}^T \gamma_{-1}) (2\pi\tau_\gamma^2)^{-1/2} \exp \left\{ -\frac{1}{2\tau_\gamma^2} (\gamma_1 - \mu_\gamma)^2 \right\} \\ \hat{\gamma}_1 &= (w_1^T w_1)^{-1} w_1^T \tilde{y} \\ \mu_\gamma &= (w_1^T w_1 + \sigma^2/c)^{-1} w_1^T \tilde{y} \\ \tau_\gamma^2 &= \sigma^2 (w_1^T w_1 + \sigma^2/c)^{-1}. \end{aligned}$$

The two integrals with respect to  $\gamma_1$  may be done analytically.

If  $\iota_1 = 0$ , then  $\gamma_1$  need not be generated. If  $\iota_1 = 1$ , then

$$p(\gamma_1 \mid y, \iota_1 = 1, \Theta) \propto p(y \mid \iota_1 = 1, \gamma_1, \Theta) p(\gamma_1, \gamma_{-1} \mid \iota),$$

with  $p(\gamma_1, \gamma_{-1} \mid \iota)$  already given in (15). This is a mixture distribution that can be sampled by an efficient accept/reject algorithm.

The details for the Gaussian prior for very similar, with the modification that the term  $\gamma_1^2 + \gamma_{-1}^T \gamma_{-1}$  does not appear in front of the normal kernel, and that drawing from the mixture distribution above may be done using constrained Gaussian distributions, without appealing to an accept/reject algorithm.

## C Proof of Theorem 2

We start with a Lemma which provides a useful result for both  $f_0 \in M_1$  and  $f_0 \in M_2$ .

LEMMA 1. If  $\hat{f}_n(x)$  is the maximum likelihood estimator of the regression spline, including both monotone and non-monotone parts, and

$$n^{-1} \sum_{i=1}^n (\hat{f}_n(x_i) - f_0(x_i))^2 \rightarrow 0 \quad \text{almost surely,} \quad (17)$$

then

$$\liminf_n \inf_{f_1} n^{-1} \sum_{i=1}^n (y_i - f_1(x_i))^2 \geq \sigma_0^2 \quad \text{almost surely,}$$

and

$$\liminf_n \inf_{f_2} n^{-1} \sum_{i=1}^n (y_i - f_2(x_i))^2 \geq \sigma_0^2 \quad \text{almost surely.}$$

**Proof.** If (17) holds then we have, from the triangular inequality, and removing almost surely from what follows, that

$$n^{-1} \sum_{i=1}^n (y_i - f_0(x_i))^2 \leq n^{-1} \sum_{i=1}^n (y_i - \hat{f}_n(x_i))^2 + n^{-1} \sum_{i=1}^n (\hat{f}_n(x_i) - f_0(x_i))^2.$$

The last term goes to 0 so in the limit

$$n^{-1} \sum_{i=1}^n (y_i - f_0(x_i))^2 \leq n^{-1} \sum_{i=1}^n (y_i - \hat{f}_n(x_i))^2.$$

But from the definition of the maximum likelihood estimator,

$$n^{-1} \sum_{i=1}^n (y_i - \hat{f}_n(x_i))^2 \leq n^{-1} \sum_{i=1}^n (y_i - f_0(x_i))^2,$$

and hence, as

$$n^{-1} \sum_{i=1}^n (y_i - f_0(x_i))^2 \rightarrow \sigma_0^2$$

it follows that

$$n^{-1} \sum_{i=1}^n (y_i - \hat{f}_n(x_i))^2 \rightarrow \sigma_0^2.$$

Then, if  $f_0$  is monotone, we have by definition

$$n^{-1} \sum_{i=1}^n (y_i - \hat{f}_n(x_i))^2 \leq n^{-1} \sum_{i=1}^n (y_i - \hat{f}_2(x_i))^2,$$

and, if  $f_0$  is non-montone, we have also by definition

$$n^{-1} \sum_{i=1}^n (y_i - \widehat{f}(x_i))^2 \leq n^{-1} \sum_{i=1}^n (y_i - \widehat{f}_1(x_i))^2,$$

completing the proof.

Conditions under which (17) holds are to be found, for example, in Geman and Hwang (1982), using sieves. So define

$$S_n = \left\{ f(x) : f(x) \text{ continuous, } f'(x) \text{ piecewise continuous, } \int |f'(x)|^2 dx < \lambda_n \right\}$$

for some increasing  $\lambda_n \uparrow \infty$ . The functions used in the present paper use the smaller sieve,  $S'_n$ , detailed in the paper, but this does not make a difference since the sieve maximum likelihood estimator from  $S_n$  actually comes from  $S'_n$ .

On pages 10 and 11 and in Theorem 2 of a Brown University technical report in 1981 by S. Geman on sieves for nonparametric estimation of densities and regressions, it is shown that  $\widehat{f}_n(x)$ , the minimizer of

$$\sum_{i=1}^n (y_i - f(x_i))^2,$$

subject to  $f \in S_n$ , is piecewise linear, with knots at the  $(x_i)$ , and if

$$\int \exp(t|y|) F_Y(dy) < \infty$$

for some  $t > 0$ , which we assume to be the case, then  $\lambda_n = O(n^{1/4-\delta})$  for some  $\delta > 0$ , is sufficient for

$$\int_0^1 \left( \widehat{f}_n(x) - f_0(x) \right)^2 Q(dx) \rightarrow 0 \quad \text{almost surely}$$

It is then easy to show, assuming  $f_0$  is bounded, that

$$\int_0^1 \left( \widehat{f}_n(x) - f_0(x) \right)^2 Q_n(dx) \rightarrow 0 \quad \text{almost surely}$$

where  $Q_n$  is the empirical distribution of the  $(x_i)$ . Hence, (17) follows.

**1.  $f_0$  is non-monotone.** Let us first assume that  $f_0(x)$ , the true mean function on  $\mathbb{X}$ , is non-monotone. That is, the true density is  $g_0(y|x, \theta_0) = N(y|f_0(x), \sigma_0^2)$ . So consider

$$B_{12} = \frac{\int \prod_{i=1}^n N(y_i|f_1(x_i), \sigma_1^2) \pi_1(d\theta_1)}{\int \prod_{i=1}^n N(y_i|f_2(x_i), \sigma_2^2) \pi_2(d\theta_2)}.$$

The denominator, with additional factor involving the true model, is given by

$$I_{n2} = \int \prod_{i=1}^n \frac{N(y_i|f_2(x_i), \sigma_2^2)}{N(y_i|f_0(x_i), \sigma_0^2)} \pi_2(d\theta_2).$$

Hence, provided the prior for  $(f_2, \sigma_2^2)$  has  $f_0$  in the Kullback–Leibler support, that is

$$\pi_2 \left\{ (f_2, \sigma_2) : \int d_K(N(\cdot|f_2(x), \sigma_2^2), N(\cdot|f_0(x), \sigma_0^2)) Q(dx) < \epsilon \right\} > 0$$

for all  $\epsilon > 0$ , then  $I_{n2} > e^{-n\tau}$  almost surely for all large  $n$ , for any  $\tau > 0$ .

The numerator can be written as

$$I_{n1} = \int \prod_{i=1}^n \frac{N(y_i|f_1(x_i), \sigma_1^2)}{N(y_i|f_0(x_i), \sigma_0^2)} \pi_1(d\theta_1)$$

and  $f_1$  belongs to the set of monotone functions. We need to show that  $I_{n1} < e^{-n\delta}$  almost surely for all large  $n$  for some  $\delta > 0$ , and we can do this with the following two conditions: the first of which is from Lemma 1,

1. It is that

$$\liminf_n \inf_{f_1} n^{-1} \sum_{i=1}^n (y_i - f_1(x_i))^2 \geq \sigma_0^2 \text{ almost surely}$$

2. If the  $x$  are sampled from  $Q$  then, for some constant  $\psi > 0$ ,

$$\sup_{f_1} \int \exp \{ -\psi (f_0(x) - f_1(x))^2 \} Q(dx) < 1.$$

First, we have

$$I_{n1} \leq \left\{ \sup_{f_1, \sigma_1^2} \prod_{i=1}^n \frac{N(y_i|f_1(x_i), \sigma_1^2)}{N(y_i|f_0(x_i), \sigma_0^2)} \right\}^{1/2} \int \left\{ \prod_{i=1}^n \frac{N(y_i|f_1(x_i), \sigma_1^2)}{N(y_i|f_0(x_i), \sigma_0^2)} \right\}^{1/2} \pi_1(d\theta_2).$$

Write these two terms as  $K_{n1}$  and  $J_{n1}$ , respectively.

If we let  $\hat{f}_1(x_i)$  be the minimizer of

$$\sum_{i=1}^n (y_i - f_1(x_i))^2,$$

then the appropriate  $\hat{\sigma}_1^2$  is given by

$$n^{-1} \sum_{i=1}^n (y_i - \hat{f}_1(x_i))^2.$$

Hence,

$$K_{n1} = \left( \frac{\sigma_0^2}{n^{-1} \sum_{i=1}^n (y_i - \widehat{f}_1(x_i))^2} \right)^{n/2} \exp \left\{ -\frac{1}{2}n + \frac{1}{2}n n^{-1} \sum_{i=1}^n (y_i - f_0(x_i))^2 / \sigma_0^2 \right\}.$$

Since

$$n^{-1} \sum_{i=1}^n (y_i - f_0(x_i))^2 / \sigma_0^2 \rightarrow 1 \quad \text{almost surely,}$$

and using condition 1., it follows that  $K_{n1} < e^{n\eta}$  almost surely for all large  $n$ , for any  $\eta > 0$ .

On the other hand, the expectation of  $J_{n1}$  is given by

$$E J_{n1} = \int \left\{ 1 - \frac{1}{2} \int d_H^2(N(\cdot|f_1(x), \sigma_1^2), N(\cdot|f_0(x), \sigma_0^2)) Q(dx) \right\}^n \pi_1(d\theta_1).$$

Now

$$\begin{aligned} \frac{1}{2} d_H^2(N(\cdot|f_1(x), \sigma_1^2), N(\cdot|f_0(x), \sigma_0^2)) &= 1 - \sqrt{\frac{2\sigma_0\sigma_1}{\sigma_1^2 + \sigma_0^2}} \exp \left\{ -\frac{(f_0(x) - f_1(x))^2}{4(\sigma_0^2 + \sigma_1^2)} \right\} \\ &\geq 1 - \exp \{ -\psi (f_0(x) - f_1(x))^2 \} \end{aligned}$$

for some constant  $\psi > 0$ , if we impose an upper bound on the prior for  $\sigma_1^2$ .

Therefore, with condition 2.,  $E J_{n1} < e^{-n\kappa}$ , for some  $\kappa > 0$ , and hence  $J_{n1} < e^{-n\phi}$  almost surely for all large  $n$ , for some  $\phi > 0$ .

Hence,  $I_{n1} < e^{n\eta} e^{-n\phi}$  almost surely for all large  $n$  for any  $\eta > 0$  and for some  $\phi > 0$ , yielding  $I_{n1} < e^{-n\delta}$  almost surely for all large  $n$  for some  $\delta > 0$ . Putting all this together

$$B_{12} \leq \exp\{n(-\delta + \tau)\}$$

almost surely for all large  $n$  for any  $\tau > 0$  and for some  $\delta > 0$ . So choose  $\tau < \delta$  to get the required result that  $B_{12} \rightarrow 0$  almost surely as  $n \rightarrow \infty$ .

Condition 2., i.e.

$$\sup_{f_1} \int \exp \{ -\psi (f_0(x) - f_1(x))^2 \} Q(dx) < 1$$

must hold since  $f_0(x)$  is a fixed non-monotone function and it is not possible for a monotone function to get arbitrarily close to it.

**2.  $f_0$  monotone.** Now let us consider the reverse case when  $f_0$  is a fixed monotone function. Following the same style of proof we have a number of considerations; amounting to the Kullback–Leibler condition, which we take as given, and,

3. It is that

$$\liminf_n \inf_{f_2} n^{-1} \sum_{i=1}^n (y_i - f_2(x_i))^2 \geq \sigma_0^2 \quad \text{almost surely}$$

4. For some constant  $\psi > 0$  and  $\rho_0 < 1$ , it is that

$$\int \left( \int \exp\{-\psi(f(x) - f_0(x))^2\} Q(dx) \right)^n \pi_2(df) < \rho_0^n. \quad (18)$$

for all large  $n$ .

Condition 3. follows from Lemma 1. For condition 4., we can split the outer integral into two parts: one with

$$\int \exp\{-\psi(f(x) - f_0(x))^2\} Q(dx) < 1 - \epsilon$$

and the other with

$$\int \exp\{-\psi(f(x) - f_0(x))^2\} Q(dx) > 1 - \epsilon$$

for a sufficiently small  $\epsilon$ . The case for the inner integral being bounded above by  $1 - \epsilon$  is clear, so we need to show that

$$\pi_2 \left\{ f_2 : \int \exp\{-\psi(f_0(x) - f_2(x))^2\} Q(dx) > 1 - \epsilon \right\} < \rho^n, \quad (19)$$

for some  $\rho < 1$ .

Suppose we have  $Q(dx)$  which takes  $n$  equi-spaced samples  $(x_1, \dots, x_n)$  and, for each  $f_2$ , define  $k$  as the number of points for which  $|f_2(x_i) - f_0(x_i)| < \delta$ . The condition now translates to

$$\pi_2(k/n > 1 - \delta^*) < \rho^n$$

for some  $\delta^* > 0$ . To see this, and letting  $z_i = \psi(f_0(x_i) - f_2(x_i))^2$ , we need to consider

$$n^{-1} \sum_{i=1}^n e^{-z_i} > 1 - \epsilon.$$

Putting  $\phi_i = 1 - e^{-z_i}$ , we need to consider

$$n^{-1} \sum_{i=1}^n \phi_i < \epsilon,$$

and  $z_i < \delta$  implies  $\phi_i < \delta$ . Now putting  $\delta = M\epsilon$ , for some  $M > 1$ , then we need at least  $k$  of the  $\phi_i$  to be less than  $\delta$  where

$$n^{-1}(n - k)M\epsilon < \epsilon.$$

Thus,  $k > n(1 - 1/M)$ .

Now  $P(k > n(1 - \delta^*))$  is bounded above by  $P(k_u > n(1 - \delta^*))$ , where  $k_u$  is the number of increasing terms for the non-monotone function. This follows since if the



derivative of  $f_0(x)$  is bounded away from 0, then  $f(x_i)$  can only be close to  $f_0(x_i)$  if the function increases in the interval immediately preceding  $x_i$ .

Hence, for some  $\rho_1$ ,  $k_u$  is binomial  $(\rho_1, n)$ , and using a normal approximation,

$$k_u/n \approx N(\rho_1, \rho_1(1 - \rho_1)/n)$$

and hence

$$P(k > n(1 - \delta^*)) < P(k_u > n(1 - \delta^*)) \approx P(z > \xi\sqrt{n})$$

for some  $\xi > 0$ , and where  $z$  is a standard normal r.v. Using the asymptotic expression for the survival function of the normal cdf, we have

$$P(k_u > n(1 - \delta^*)) \approx \frac{c_1}{\sqrt{n}} \exp(-c_2 n)$$

for positive constants  $c_1$  and  $c_2$ . Hence, (19) holds true, and so (18) holds true.

## References

- J. O. Berger and L. Pericchi. Objective Bayesian methods for model selection: introduction and comparison. In *Model Selection*, volume 38 of *Institute of Mathematical Statistics Lecture Notes – Monograph Series*, pages 135–207. Beachwood, 2001.
- A. Bowman, M. Jones, and I. Gijbels. Testing monotonicity of regression. *Journal of Computational and Graphical Statistics*, 7(4):489–500, 1998.
- B. Cai and D. B. Dunson. Bayesian multivariate isotonic regression splines. *Journal of the American Statistical Association*, 102(480):1158–71, 2007.
- C. M. Carvalho, M. S. Johannes, H. F. Lopes, and N. G. Polson. Particle learning and smoothing. *Statistical Science*, 25(1):88–106, 2010.
- P. Chigansky and F. Klebaner. Distribution of the Brownian motion on its way to hitting zero. *Electronic Communications in Probability*, 13:641–8, 2008.
- M. Clyde and E. I. George. Model uncertainty. *Statistical Science*, 19(1):81–94, 2004.
- M. A. Clyde and R. Wolpert. Nonparametric function estimation using overcomplete dictionaries. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Proceedings of the 8th Valencia World Meeting on Bayesian Statistics*, pages 91–114. Oxford University Press, 2007.
- I. Dimatteo, C. R. Genovese, and R. E. Kass. Bayesian curvefitting with freeknot splines. *Biometrika*, 88(4):1055–71, 2001.
- D. Dunson. Bayesian semiparametric isotonic regression for count data. *Journal of the American Statistical Association*, 100(470):618–27, 2005.
- S. Geman and C. Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics*, 10(2):401–14, 1982.

- E. I. George and D. P. Foster. Calibration and empirical Bayes variable selection. *Biometrika*, 87(4):731–747, 2000.
- F. J. Girón, E. Moreno, G. Casella, and M. L. Martínez. Consistency of objective Bayes factors for nonnested linear models and increasing model dimension. *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales (Serie A: Matemáticas)*, 104(1):57–67, 2010.
- L. Hannah and D. Dunson. Bayesian nonparametric multivariate convex regression. Technical report, Duke University Department of Statistical Science, 2012a.
- L. Hannah and D. Dunson. Multivariate convex regression with adaptive partitioning. Technical report, Duke University Department of Statistical Science, 2012b.
- C. Holmes and N. Heard. Generalized monotonic regression using random change points. *Statistics in Medicine*, 22:623–38, 2003.
- V. E. Johnson and D. Rossell. On the use of non-local prior densities in bayesian hypothesis tests. *Journal of the Royal Statistical Society (Series B)*, 72(2):143–70, 2010.
- J. S. Liu and M. West. Combined parameters and state estimation in simulation-based filtering. In A. Doucet, N. de Freitas, and N. Gordon, editors, *Sequential Monte Carlo Methods in Practice*. Springer, 2001.
- B. Neelon and D. Dunson. Bayesian isotonic regression and trend analysis. *Biometrics*, 60: 398–406, 2004.
- A. Panagiotelis and M. Smith. Bayesian identification, selection and estimation of semi-parametric functions in high-dimensional additive models. *Journal of Econometrics*, 143(2):291–316, 2008.
- L. Schwartz. On bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 4(1):10–26, 1965.
- J. G. Scott. Nonparametric Bayesian multiple testing for longitudinal performance stratification. *The Annals of Applied Statistics*, 3(4):1655–74, 2009.
- T. S. Shively, T. Sager, and S. G. Walker. A Bayesian approach to nonparametric monotone function estimation. *Journal of the Royal Statistical Society (Series B)*, 71:159–75, 2009.
- T. S. Shively, S. G. Walker, and P. Damien. Nonparametric function estimation subject to monotonicity, convexity, and other shape constraints. *Journal of Econometrics*, 161: 166–81, 2011.
- M. Smith and R. Kohn. Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75:317–43, 1996.
- S. G. Walker and N. L. Hjort. On bayesian consistency. *Journal of the Royal Statistical Society (Series B)*, 63:811–21, 2002.
- J. X. Zheng. A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics*, 75(263–89), 1996.